Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 385 – 394

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Robust Temporal Registration Scheme for Video Copies Using Visual-Audio Features

Dr. R. Roopalakshmi[a,*], Revanur Venkatesh[a], K.M. Rahul[a]

[a]*R.V. College of Engineering, Bangalore, Karnataka 560059, India.*

Abstract

Followed by video copy detection, temporal frame alignments of the copied video with the master contents is essential in numerous forensic applications such as, computation of geometric distortions and estimation of pirate location in a theater during illegal camcorder captures. State-of-the-art temporal video copy registration methods are exploiting only visual features of videos, while no effort is made to employ audio signatures. Furthermore, existing studies are primarily focusing on the alignment of watermarked videos, while very few efforts are made towards non-watermarked videos. To solve these issues, this paper presents a robust temporal registration scheme by utilizing visual-audio fingerprints, which consists of two stages: First, the video sequence is compactly represented using 1-D motion and acoustic profiles; Second, accurate frame-to-frame matches are computed using sliding window based dynamic programming technique. Experiments on TRECVID-2008 & 2009 datasets, prove the efficiency and effectiveness of the proposed framework compared to the reference methods against a wide range of video editing and transformations.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: Video copy; Temporal Registration; Motion Vectors; MFCC; Dynamic Programming.

1. Introduction

A *video copy/pirate video* is a distorted video sequence derived from the master video by applying different video editing and transformations such as noise and caption insertions. In this paper, we define the term "*registration*", which represents a way of mapping master and pirate video contents with an objective to compute accurate frame-to-frame alignments. With the exponential growth of multimedia streaming and online sharing activities, a huge number of pirated versions of movies are available before their official release and cause a great loss to motion picture industry. For instance, according to Canadian Motion Picture Distributors Association (CMPDA-2010) report[1], 133 million pirated movies are watched in Canada in 2010. This survey also indicates a total loss of C$413 million to Canadian economy due to Internet based digital piracy. Hence rigorous forensic analysis frameworks and countermeasures are required to restrict video piracy.

* Corresponding author. Tel.: +91 9972246013 ; fax: +0-080-67178011.
E-mail address: roopalakshmir@rvce.edu.in

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.353

Combating camcorder piracy requires copy detection as the first step, which attempts to find out the best matching master video for a given query clip. Digital watermarking and Content-Based video Copy Detection (CBCD) are the two standard techniques used to detect video copies[2]. CBCD techniques employ content-based features of the media to detect duplicate videos[3]; hence they are most successful compared to digital watermarking[3,4]. However, existing CBCD methods do not address temporal frame alignments of the copied video with the master sequence, because their ultimate goal is to detect video copies by comparing the perceptual similarity between the two video sequences. On the other hand, in case of illegal Camcorder captures in a theater, significant misalignments exist between the master and pirate video sequences, which could be temporal, geometric or the combination of both. Due to these reasons, temporal registration of the pirate video with the master contents is very much essential, for a number of applications such as geometric distortions estimation and pirate position identification in a theater[5]. For instance, temporal frame alignments are successfully employed, in order to obtain the accurate geometric frame matches of the pirate video with the master content[6]. Further, temporal frame-to-frame alignments are utilized in[7], which estimates the geometric distortions present in the duplicate video in terms of projection matrix. Furthermore, in order to detect the forensic watermarks embedded in copied clips, we need to first register the copied sequence with the original video[8].

State-of-the-art video copy registration techniques are utilizing only visual features of videos for obtaining accurate frame alignments of two video sequences. For example, Baudry et al.[9] used wavelets-based video signatures and dynamic programming method for decoding the embedded watermarks, in order to achieve temporal registration of pirate and master contents. Though, this method guarantees accurate alignments, its computational cost and memory usage is high. Delannay et al.[10] proposed key frames based registration technique to temporally align the two video sequences, which fails for high motion activity frames. Baudry et al.[11] used both the global and local fingerprints for registering video sequences. However, their method scores poor results for low motion frames and complex transformations such as letter-box insertion and subtitles. In[12], Chupeau et al. used color histograms as video signatures and matched two video contents using dynamic programming method, which fails towards region-based transformations. Hui Cheng[13] proposed a temporal registration algorithm to match two video sequences using dynamic programming method, which is severely affected by transformations such as noise addition. Cheng and Isnardi[14] developed a spatial, temporal and histogram based registration algorithm to detect forensic watermarks, which focuses only on watermarked contents. In[15], author discussed and compared three different registration algorithms that are specifically designed for detecting embedded watermarks in digital cinema applications.

To summarize, existing video copy temporal registration schemes are exploiting only visual features of videos, while no effort is made to employ audio signatures. However, audio content is an indispensable and essential information source of a video sequence. Further, in most of the illegal Camcorder captures, the audio content of an illegal video is less affected compared its visual part[16]. From another perspective, state-of-the art registration schemes are focusing only on the alignment of watermarked documents. However, all video contents are not watermarked. Therefore, promising temporal registration schemes making use of visual as well as acoustic features are required, which can be used even in the absence of forensic watermarks.

Contributions. This paper introduces a new temporal registration framework that utilizes visual-audio fingerprints for obtaining frame-to-frame alignments of the copied video with the master content. Specifically, 1-D motion profile extracted from motion vector magnitudes and 1-D acoustic profile derived from MFCCs are employed in this temporal registration task in order to obtain the accurate temporal frame alignments of two video sequences. More specifically, first we present a *candidate segment selection* algorithm, for selecting the most similar segment of the master sequence using sliding window based dynamic programming technique, which noticeably reduces the frame matching cost. Further, the proposed framework, introduces a frame matching scheme exploiting multimodal features for achieving temporal frame matches, which significantly reduces false frame matches.

2. Proposed Method

2.1. Proposed Temporal Registration Framework

Fig. 1, shows the proposed framework for temporally registering the duplicate and master video contents, which consists of two steps: First, temporal signatures are derived from motion and acoustic features of two video files;

Second, the resultant visual-audio fingerprints of master and query segments are matched separately using dynamic programming technique, in order to obtain accurate frame-to-frame matches. More specifically, when a query clip is given, we divide the master sequence into non overlapping segments of size equal to length of the query frames. Then we perform segment-wise scanning of the master sequence using a sliding window of length equal to query clip. The similarity between the query clip and the windowed segment is computed based upon their 1-D signatures derived from motion vector magnitudes and MFCCs. The windowed segment with minimum dissimilarity score (score below a predefined threshold) is denoted as the *candidate segment*, and it is further analyzed using dynamic programming technique to determine the exact frame-to-frame alignments of two video sequences.

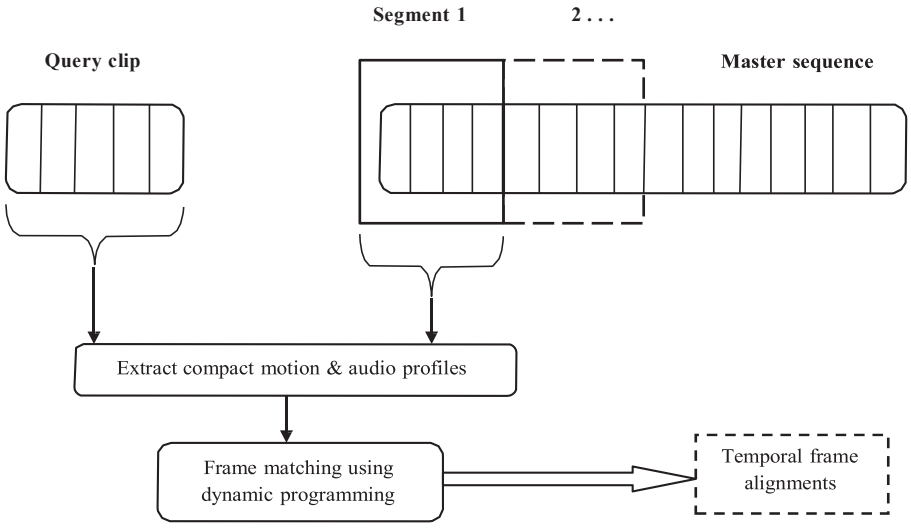

Fig. 1. The Proposed Temporal Registration Framework

2.2. Temporal Frame Alignments

Comparing multi-dimensional signatures of two video contents is a tedious process because of their huge size. In order to handle this issue, in the proposed system, a video sequence is compactly represented using 1-D temporal signatures, which are not only easy to manipulate but also robust against various video distortions. The 1-D temporal signatures including motion and acoustic profiles of two video sequences are extracted as described below.

2.2.1. Compact Motion Profile Extraction

Motion vectors are one of the popular temporal features used in various video analysis applications such as indexing, retrieval and video characterization [17 18]. However, in the CBCD literature, motion features are considered as poor descriptors due to these reasons: a) Motion vectors are almost equal to zero values, when they are captured at normal frame rates (25-30 fps); b) Raw motion vectors are noisy and require huge amount of information for describing the motion content. In order to solve this discrepancies, we captured motion vectors at a lower frame rate (4 or 5fps), which are descriptive enough to represent the given video content.

On the other hand, though Motion vector magnitudes provide better temporal information of video contents; yet they fail to describe the complete spatial content. *If the spatial-distribution of motion content in a given frame is also utilized along with the temporal motion description, then it is possible to generate a robust motion profile of video contents*. Based on this aspect, we employed spatio-temporal motion descriptors in the proposed framework for registering the given two video contents. Precisely, the 1-D motion profile is computed as the sum of differences between region-wise motion vector magnitudes of consecutive frames. More precisely, we segmented the frame into

$n \times n$ regions and computed average motion vector magnitude of regions. The average motion vector magnitude (AMV) of k-th region R_K is given by,

$$R_k(AMV) = \frac{1}{MN}\sum_{i=1}^{M}.\sum_{j=1}^{N} mv(i,j) \quad (1)$$

where $mv(i, j)$ represents motion vector magnitude of (i, j)-th block of region R_k, such that $k = \{1, 2, 3, ..., n \times n\}$ and MN is the region size. The segmentation of a frame into $n \times n$ regions, plays a significant role in predicting spatial motion content in a given frame. Smaller values of n may leave important semantic content, whereas larger values of n increase the computational process. In order to solve this problem, we tested our data sets with different n values ranging from 2 to 5, while maximum accuracy is achieved when n=3. Thus we computed spatial motion distribution by segmenting frames into 3×3 regions.

2.2.2. Compact Acoustic Profile Extraction

MFCCs are highly robust and discriminative features, thus they are widely used in video indexing and segmentation methods[19]. The MFCCs are computed using the discrete cosine transform of the log amplitude Mel-frequency spectrum. Since MFCCs consider the nonlinear property of the human hearing system with respect to different frequencies, they are popularly used in automatic speech recognition systems[19].

In the proposed scheme, first the audio signal is down sampled to 22050 Hz and segmented into 11.60 ms windows with an overlap factor of 86% using Hamming window function[20]. From the resultant spectrum, first 13 MFCCs are calculated, to capture short term acoustic features of video files. However, the raw MFCCs are huge and may contain redundant data, hence it not desirable to perform any computations. To solve this problem, we utilize only MFCC variance values to generate 1-D acoustic profile of video contents, instead of all 13 MFCCs of individual video frames.

2.2.3. Sequence Matching Using Dynamic Programming

Dynamic programming is an efficient recursive technique, which is popularly used in sequence-to-sequence alignment and comparison methods[21]. The proposed framework uses dynamic programming technique to compute frame-to-frame matches, which includes the following two steps:

1. Computing minimum score matrix:
   In order to specify the optimal alignment between two sequences, first 2-D score matrix computation is needed. An element $S(i, j)$ of score matrix S gives minimum matching cost to match subsequences [0,1,.., i] with [0,1,.., j]. The element $S(i, j)$ is recursively computed as,

$$S(i,j) = Min \begin{cases} S(i-1,j-1) \\ S(i,j-1) + W_h \\ S(i-1,j) + W_v \end{cases} + \quad Dist(i,j) \quad (2)$$

   where W_h, W_v are penalties associated with horizontal and vertical directions and $Dist(i,j)$ is the difference between two feature sequences associated with elements i and j.
2. Determining optimal alignment path: optimal frame-to-frame matches are computed, by performing a trace- back step starting from the diagonal element to determine the actual alignments.

2.2.4. Sliding Window Based Dynamic Programming

The computational complexity of dynamic programming method to match two sequences of size M and N is $O(MN)$ and memory usage is also $O(MN)$. Hence if sequence size increases, the complexity of the algorithm also increases. In order to overcome this problem, the proposed registration framework performs frame matching between the query clip and the candidate segment instead of the total master sequence. The candidate segment selection algorithm is detailed in Fig. 2. Once candidate segment is selected, then dynamic programming technique is used to compute frame-to-frame matching between query and candidate segments. Precisely, the feature sequences of query and candidate segments are separately matched using dynamic programming method to get accurate frame matches, which is detailed as follows.

a) Segment the master sequence into non-overlapping blocks of length equal to the query clip.
b) For each segment, 1-D motion and audio profiles are extracted using the procedures explained in sections 3.2.1 and 3.2.2.
c) Let the master sequence $MS = \{S_i | i = 1, 2, ..., m\}$, where S_i is the i^{th} segment and m is the total segments of the master video. Thus, $S_1 = \{mf_k \cup af_r \mid k = 1, 2, ..., n \text{ and } r = 1, 2, ..., p\}$, where mf_k is the k^{th} motion based signature of S_1 and af_r is the r^{th}h MFCCs based signature of S_1.
d) Let the query clip QS is described using 1-D motion and acoustic signatures, such that $QS = \{qmf_k \cup qaf_r \mid k = 1, 2, ..., n \text{ and } r = 1, 2, ..., p\}$, where qmf_k is the k^{th} motion based signature of QS and qaf_r is the r^{th} MFCCs based signature of QS.
e) The similarity measure (Seg_{sim})between query clip and k^{th} segment of master sequence is computed using Manhattan distance as follows,

$$Seg_{Sim}(S_k, \text{QS}) = \sum_{i=1}^{n} \left| mf_i^k - qmf_i \right| + \sum_{j=1}^{p} \left| af_j^k - qaf_j \right| \quad (3)$$

where S_k is the k^{th} segment of MS and n, p indicate the size of motion and MFCCs based features of video contents respectively.
f) A master segment with least S_{Sim} value is selected as the candidate segment of master sequence for further comparison.

Fig. 2. Candidate Segment Selection Algorithm

2.3. Frame Matching Using Visual-Audio Fingerprints

2.3.1. Motion Features Based Frame Matching

Let CS be a candidate segment of master sequence and QS be a query segment. Let mf_k and qmf_k are the motion profiles of segments CS and QS, such that $k = \{1, 2, ..., n\}$. In order to compute difference between the motion profiles of CS and QS, Comparative Manhattan distance measure is used as given by,

$$Dist_{Motion}\big(C(i)Q(i)\big) = \frac{|(mf_{(i)} - qmf_{(i)})|}{|(mf_{(i)})| + |(qmf_{(i)})|} \quad (4)$$

where $i = \{1, 2, ..., n\}$ and n are the total motion signatures of two video sequences. Then score matrix S for matching motion features is computed using equations (3) and (4). Finally optimal alignment path is determined and Frame Matches-1 (FM_1) based on motion signatures are calculated and stored in FM_1.

2.3.2. MFCCs Based Frame Matching

Let $C = \{ af_k, |k = 1, 2, ..., p\}$ and $Q = \{ qaf_k, |k = 1, 2, ..., p\}$ are the MFCCs based signatures of segments CS and QS, where p indicates the size of MFCC signatures of video files. Then the similarity between audio signatures of CS and QS is computed using Squared Euclidean distance as follows,

$$Dist_{Audio}\big(C(j), Q(j)\big) = |(af_{(j)} - qaf_{(j)})^2| \quad (5)$$

where $j = \{1, 2, ..., p\}$. Then the score matrix S for matching audio features is calculated equations (3) and (4). Finally optimal alignment path is determined and Frame Matches-2 (FM_2) based on MFCCs are computed and stored in FM_2.

2.3.3. Decision Fusion

Final frame matches ($Final_{FM}$) between query and candidate segments are computed as given by,

$$Final_{FM} = |FM_1 \cap FM_2|. \quad (6)$$

In the proposed registration framework, only frames with similar visual and audio signatures are mapped and hence it significantly reduces false frame matches.

3. Experimental Setup and Results

3.1. Master Video Database and Query Construction

The proposed method is evaluated on TRECVID-2008, 2009 Sound & Vision data sets[22], which are used as benchmark data sets for CBCD task. The video database includes totally 250 hours of video (100hrs of 2008 data + 150hrs of 2009 data) covering a wide variety of content. All the video clips are converted into uniform format: 352×288 pixels and 5 frames/sec. Table 1 lists different types of video transformations used in proposed registration task such as geometric, temporal, filtering, audio and combined transformations. 50 video clips are randomly selected from the master video database and duration of these clips vary between 20-52 seconds. Seventeen types of video transformations listed in Table 1 are applied to these video clips to generate total query video set. The resulting 850(50 ×17) video sequences are used as query clips for the proposed registration task. The snapshot of GUI of the proposed temporal registration framework is shown in Fig.3.

Table 1. List of video transformations used in the proposed registration scheme

Category	Type	Description
Geometric	Rotation	Rotating by 45° to 95°
	Cropping	Crop top & bottom regions by 20% each
	Flipping	Horizontal flip by 120°-180°
Temporal	Fast forward	Double the video speed
	Slow motion	Halve the video speed
Pattern	Pattern insertion	Insert text pattern into selected frames
	Picture-in-picture	Insert smaller resolution picture into selected frames
	Moving caption	Insert moving titles into entire video
Filtering	Blurring	Blurring by 28%
	Noise addition	Add 15% gaussian noise
	Contrast change	Increase contrast by 20%
Scaling	Zooming in	Zoom in to the frame by 13%
	Resolution change	Change frame resolution to 150×120 pixels
Audio	Mp3 compression	Change audio file format
	Single band comp.	Compress only specific frequency band
	Multi band comp.	Compress different frequency bands independently
Combined	3 combined	Cropping by 18%, 20% of noise & moving caption

3.2. Overview of Evaluated Methods

We implemented the following six methods for evaluating performance:

(1) The motion features based matching without sliding window(abbreviated as ”MV”);
(2) The motion features based matching with sliding window(”MV+SW”);
(3) MFCCs based matching without sliding window(”MFCC”);
(4) MFCCs based matching with sliding window(”MFCC+SW”);
(5) Chupeau et al.’s method[12](”CH”);
(6) The combination of motion and MFCCs with sliding window (methods(2) and (4))(”ALL”);

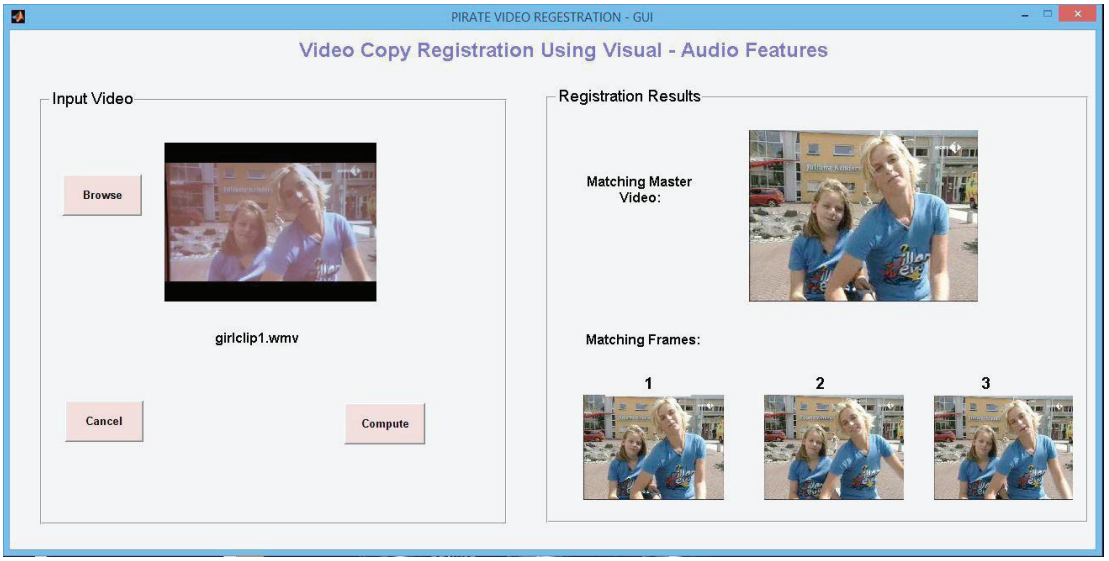

Fig. 3. Snapshot of GUI of Proposed Video Copy Registration System

Our methods [methods(1)-(4) and (6)] evaluated different combinations of proposed techniques. Methods(1) and (3) used different video signatures (namely motion vectors and MFCCs) to perform temporal registration of two video contents. We implemented methods(2) and (4) to see the effect of sliding window technique for the proposed frame registration task. In method (1) 1-D motion features of query clip are matched with the 1-D motion profile of the entire master sequence (i.e. query clip is matched with all segments of the master sequence). In method(2), we used sliding window mechanism to align query motion features with that of candidate segment. The selection of candidate segment of master sequence is implemented using the procedure explained in section 2.2.4. 1-D MFCC signatures of query clip are mapped with the acoustic profile of the complete master sequence in method(3). In method(4), sliding window approach is utilized to match query MFCC features with the corresponding features of candidate segment.

Chupeau et al.'s method (method(5)) uses color histograms for calculating frame-to-frame correspondences between query and master video sequences. It is implemented as follows: color histograms of size 512 bins are extracted from consecutive video frames. A sequence of distances (Euclidean distance) between color histograms of successive frames are utilized as temporal fingerprints of video files. In method(6), both the motion and MFCC signatures of query clip are matched separately with the corresponding features of the candidate segment in order to get accurate frame-to-frame alignments.

3.3. Registration Accuracy Comparison

In the following subsections, the registration results of six compared methods in terms of various video (including visual and audio)transformations are discussed.

3.3.1. Geometric and Scaling Transformations

Table 2 lists the registration accuracy of six compared methods for geometric and scaling transformations. This category includes rotation, cropping, flipping, zooming in and resolution change. Methods(3), (4) and (6) generally perform well, when compared to methods(1), (2) and (5). There is a slight improvement in the registration accuracy (by 3.2%) of method(2), compared to that of method(1). The reason for this improvement is, the usage of sliding window scheme significantly reduces false positive rate. Method(4) slightly improves the accuracy compared to that of method(3), because of the incorporation of sliding window scheme which reduces false positive rate.

Method(6) performs better for all six transformations by improving registration accuracy (up to 41.9%) compared to the reference methods. Integration of both motion and audio features for frame registration is the exact reason for

Table 2. Perfectly registered frames (in %) for geometric and scaling transformations

Transformations		MV	MV+SW	MFCC	MFCC	CH	ALL
Category	Type	(1)	(2)	(3)	+SW(4)	(5)	(6)
Geometric	Rotation	54.45	58.67	91.82	91.82	58.83	93.29
	Cropping	53.59	59.31	90.71	90.85	49.62	92.71
	Flipping	46.72	50.77	90.63	91.69	50.07	94.68
Scaling	Zooming in	52.61	52.99	91.56	92.18	48.85	92.49
	Resolution change	57.61	59.45	89.57	90.37	49.26	93.18

Table 3. Perfectly registered frames (in %) for temporal and caption transformations

Transformations		MV	MV+SW	MFCC	MFCC	CH	ALL
Category	Type	(1)	(2)	(3)	+SW(4)	(5)	(6)
Temporal	Slow motion	53.08	54.96	65.71	66.77	51.63	90.83
	Fast forward	45.15	48.27	62.93	62.93	50.74	88.75
Caption	Pattern insertion	61.57	65.15	90.53	90.62	45.71	91.03
	Picture-in-picture	49.57	53.64	91.78	92.46	48.94	92.85
	Moving caption	53.02	55.25	89.96	90.05	46.68	91.94

this improved performance. On the other hand, Chupeau et al.'s method (method (5)) yields poor results for flipping and zooming transformations. The reason for the poor performance of method(5) is, the limited capabilities of color histograms against region-based transformations.

3.3.2. Temporal and Caption Transformations

Table 3 shows the registration accuracy of six compared methods for temporal and caption based transformations. This category includes slow motion, fast forward, pattern insertion, picture-in-picture and moving caption transformations. Method(5) gives poor results for caption based transformations. This is because, inserting text patterns substantially changes the histogram based signatures. But our methods using MFCC features (methods(3), (4) and (6)) are less affected by this category of transformations.

We observe that the method(6), which combines MFCC and motion features for frame matching significantly improves registration accuracy by 40-45% for all five transformations listed in Table 3. For fast forward transformation, method (6) performs well and significantly increases registration accuracy by 43.6% when compared to the reference methods.

3.3.3. Audio and Filtering Transformations

Table 4 shows the registration results of six compared methods for audio and filtering transformations. Audio transformation category includes mp3, single band and multi band compressions. Filtering category includes blurring, noise addition and contrast change transformations. The registration accuracy of only MFCC based methods (method(3) and (4)) degrade slightly for audio transformations. This is because the spectral coefficients are much affected by single and multi band compressions. The motion features are much affected by filtering attacks such as noise addition, which in turn reduces the registration rates of methods(1) and (2) for filtering transformations. The Table 4 results demonstrate the improved performance of method(6)(up to 33.3%) for all seven transformations compared to other evaluated methods.

3.4. Computation Cost Comparison

To evaluate the proposed method, we implemented the code in MATLAB using a PC with 2.8GHz CPU and 3 GB RAM. The total computational cost of all six methods including signatures extraction and matching are shown in Table 5. They are measured using 24s query clip and 2944s master sequence for temporal registration. For example, methods(1)-(6) take 234.63s, 177.58s, 150.36s, 103.54s, 182.09s and 173.06s respectively to register a query clip

Table 4. Perfectly registered frames (in %) for audio, filtering and combined transformations

Transformations		MV	MV+SW	MFCC	MFCC	CH	ALL
Category	Type	(1)	(2)	(3)	+SW(4)	(5)	(6)
Audio	mp3 comp.	75.64	75.94	56.66	57.64	60.23	90.46
	Single band comp.	78.24	79.24	61.16	61.16	62.82	92.38
	Multi band comp.	76.06	76.06	61.63	61.97	61.56	91.57
Filtering	Blurring	57.70	59.31	78.59	79.38	62.77	90.34
	Noise addition	52.91	56.18	85.34	85.98	56.98	92.49
	Contrast change	62.74	59.88	84.74	82.62	51.35	91.37
Combined	3 combined	41.68	45.16	80.67	82.56	42.85	89.56

Table 5. Comparison of computational cost

Computational Cost	MV	MV+SW	MFCC	MFCC	CH	ALL
	(1)	(2)	(3)	+SW(4)	(5)	(6)
Signature extraction	176.95	175.57	103.98	102.49	156.41	171.39
Signature matching	57.68	2.02	46.38	1.06	25.68	1.67
Total cost	234.63	177.58	150.36	103.54	182.09	173.06

of duration 24s with the master sequence. The signature matching cost of method (2) is reduced drastically (nearly 96.5%) when compared to that of method (1). The reason behind this drastic reduction is, in method (2) only the candidate segment motion features are matched with the query clip features using sliding window scheme. Thus method (2) reduces the computational time by 25% compared to method (1).

There is a huge reduction (nearly 97.8%) in the fingerprint matching cost of method (4), when compared to that of method (3). The reason for this drastic reduction is, in method (4) the MFCC features of query clip are aligned only with that of candidate segment instead of the entire master sequence. Hence, method (4) reduces the computational cost by 34% when compared to method(3). The total computational cost of method (6) is slightly high compared to methods (2)-(4). Although method (4) is the most cost effective method, but its registration results are poor for audio transformations, when compared to that of method (6). Thus results prove that, method (6) significantly improves detection accuracy by 25.6% and widens the coverage to more number of transformations at the cost of slight increase in computational time.

Summary . The experimental results demonstrate that the proposed methods improve the registration accuracy. Integration of visual and acoustic features provide accurate temporal registration with reasonable robustness against wide variety of video transformations. Method (6), which integrates all proposed techniques, consistently provides better performance for all seven categories of video transformations.

4. Conclusion

In this article, we proposed a novel temporal registration scheme using visual and audio signatures. Sliding window based dynamic programming method is used to obtain accurate frame-to-frame alignments. The registration results prove that the proposed method improves accuracy by 15-42% compared to the reference methods. The results also demonstrate that the proposed method is cost effective by supporting a drastic reduction (up to 95%) in the feature matching cost. Our future work will be targeted at,

* To improve the registration accuracy of proposed method by using keypoint based features.
* To parallelize the fingerprint extraction process, which may reduce the computational complexity of proposed system.

* To enhance the robustness of proposed scheme against transformations such as camcording and the complex ones.

References

1. "Economic consequences of movie piracy", CMPDA- Feb 2011 report. http://www.mpa-canada.org/press/IPSOS-OXFORD-ECONOMICS-Report_February-17-2011.pdf
2. S.Wei, Y.Zhao, C.Zhu, C.Xu, and Z.Zhu, "Frame Fusion for Video Copy Detection", IEEE Trans. Circuits Sys. for Video Tech, vol.21, 15–28, 2011.
3. A.Sarkar, V.Singh, P.Ghosh, B.S.Manjunath, and A.Singh, "Efficient and Robust Detection of Duplicate Videos in a Large Database", IEEE Trans. Circuits & Sys. for Video Tech., 870-885, vol.20, no.6, 2010.
4. C.Y.Chiu and H.M.Wang, "Time-Series Linear Search for Video Copies Based on Compact Signature Manipulation and Containment Relation Modeling", IEEE Trans. Circuits & Sys. for Video Tech., 1603-1613, vol.20, no.11, 2010.
5. B. Chupeau, "In-theater piracy: finding where the pirate was", in Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, vol. 6819 of Proc. of SPIE, San Jose, CA, USA, 2008.
6. R.Roopalakshmi and G.R.M.Reddy, "A Novel Spatio-Temporal Registration Framework for Video Copy Localization based on Multimodal Features", Elsevier Signal Processing Journal, Vol. 93, no. 8, Pages 2339-2351, Aug'2013. Available: http://dx.doi.org/10.1016/j.sigpro.2012.06.004
7. R.Roopalakshmi and G.R.M.Reddy, "A framework for estimating geometric distortions in video copies based on visual-audio fingerprints", Springer Signal, Image and Video Processing (SIViP) Journal, Vol.7, no. 1, Jan'2013. Available: http://link.springer.com/article/10.1007/s11760-013-0424-7.2013
8. D.Delannay, F.Delaigle, H.Demarty and M.Barlaud, "Compensation of Geometrical deformations for Watermark Extraction in Digital Cinema Applications", in proc. of SPIE Electronic Imaging 2001, Security and Watermarking of Multimedia Content III, vol.4314, 149-157, 2001.
9. S.Baudry, Bertrand Chupeau and F.Lefèvreabc, "Adaptive Video Fingerprints for Accurate Temporal Registration", in proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010), pp. 1786–1789,2010.
10. D. Delannay, C. de Roover and B. Macq, "Temporal alignment of video sequences for watermarking", IS&T/SPIE's 15th Annual Symp. on Elect. Imaging, California, USA, Proc. Vol. 5020, pp. 481-492, January 2003.
11. S.Baudry, B.Chupeau and F. Lefèvreabc, "A framework for video forensics based on local and temporal fingerprints", in proc. of IEEE International Conference on Image Processing (ICIP 2009), pp. 2889–2892, 2009.
12. B. Chupeau, L. Oisel, and P. Jouet, "Temporal video registration for watermark detection", in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 157-160, Toulouse, France, 2006.
13. Hui Cheng, "Temporal Registration of Video Sequences", in proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003),Hong Kong, China, pp. 489–492,2003.
14. H. Cheng and M.A. Isnardi, "Spatial, temporal and histogram video registration for digital watermark detection", in proc. of International Conf. on Image Processing (ICIP 2003), Barcelona, Spain, pp. 735–738, 2003.
15. Hui Cheng, "A Review of Video Registration Methods for Watermark Detection in Digital Cinema Applications", in proc. of ISCAS 2004, pp. 704–707.
16. A.Saracoğlu, E.Esen, T.K.Ateş, B.O.Acar, Zubari, E.C.Ozan, E. özalp, A.A.Alatan, and T. Çiloglu, "Content Based Copy Detection with Coarse Audio-Visual Fingerprints", in proc. of 7th Int. Workshop on Content-Based Multimedia Indexing (CBMI),213-218, 2009.
17. A. Divakaran, R. Regunathan, and K. A. Peker, "Video summarization using descriptors of motion activity: A motion activity based approach to key-frame extraction from video shots," Journal of Elect. Imaging, vol. 10, pp. 909-916,2001.
18. Sofia Tsekeridou and Ioannis Pitas, "Content-Based Video Parsing and Indexing Based on AudioVisual Interaction", in proc. of IEEE Trans. on Circuits & Sys. for Video Tech., Vol. 11, No.4,2001.
19. Yao Wang, Zhu Liu, and Jin-Cheng Huang, "Multimedia Content Analysis", in proc. of IEEE Signal Processing Magazine, 12-36, 2000.
20. R.Roopalakshmi and G.Ram Mohana Reddy, "A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA", in proc. of ANT-2011 in Elsevier Procedia Computer Science Journal, 2011. doi:10.1016/j.procs.2011.07.021
21. D. Sankoff, "The early introduction of dynamic programming into computational biology", Bioinformatics, 16, 1, 4147, 2000.
22. TRECVID 2010 Guidelines [Online]. Available: http://www.nlpir.nist.gov/projects/tv2010/tv2010.html